# Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

**INSTRUCTIONS**

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

## GRANTEE SUBMISSION REQUIRED FIELDS

**Title of article, paper, or other content**

Machine Learning for Causal Inference

**All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.**

| Last Name, First Name | Academic/Organizational Affiliation | ORCID ID |
|---|---|---|
| Hill, Jennifer | New York University | 0000-0003-4983-2206 |
| Perrett, George | New York University | |
| Dorie, Vincent | Code for America | 0000-0002-9576-3064 |
| | | |
| | | |
| | | |

**Publication/Completion Date—**(if *In Press,* enter year accepted or completed) 2023

**Check type of content being submitted and complete one of the following in the box below:**

- ☐ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☒ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Handbook of Multivariate Matching and Weighting for Causal Inference, 416-443, Chapman & Hall/CRC, Boca Raton, FL

**DOI or URL to published work** (if available)

**Acknowledgement of Funding**— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** Institute of Education Sciences through **[Grant number]** R305D200019 to **Institution]** New York University .The opinions expressed are those of the authors and do not represent views of the **[Office name]** Institute of Education Sciences or the U.S. Department of Education.

# 20

# *Machine Learning for Causal Inference*

**Jennifer Hill, George Perrett, Vincent Dorie**

*Hill, Perrett, and Dorie*

## CONTENTS



DOI: 10.1201/9781003102670-20

## 20.1 Introduction

Estimation of causal effects requires making comparisons across groups of observations exposed and not exposed to a a treatment or cause (intervention, program, drug, etc). To interpret differences between groups causally we need to ensure that they have been constructed in such a way that the comparisons are "fair." This can be accomplished though design, for instance, by allocating treatments to individuals randomly. However, more often researchers have access to observational data and are thus in the position of trying to create fair comparisons through post-hoc data restructuring or modeling. Many chapters in this book focus on the former approach (data restructuring). This chapter will focus on the latter (modeling) to illuminate what can be gained from such an approach. We illustrate the case for modeling the relationship between outcomes, covariates, and a treatment to estimate causal effects using a Bayesian machine learning algorithm known as Bayesian Additive Regression Trees (BART) [1–3].

## 20.2 Causal Foundations

This section introduces the building blocks necessary to understand what causal quantities represent conceptually and why they are so difficult to estimate empirically.

### 20.2.1 Fair comparisons

At a basic level, causal inference methods require fair comparisons. For instance, suppose we want to understand the effect of a nutritional intervention on cholesterol among women with high cholesterol levels. To do so, we have access to data on 200 women and 100 of them participated in the program. All of them had high cholesterol before the date that program started. If we simply compare average cholesterol levels between those who participated in the intervention and those who didn't one year after program onset, would it make sense to think of differences in average cholesterol levels as the causal effect of the program? After all, the individuals who participated in the intervention might have been different from those who didn't in any of a variety of ways even before the program began. For example, it might be that those who participated had higher levels of cholesterol on average. Or maybe they were more motivated to make lifestyle changes to lower their cholesterol. If differences like this exist, then it would be unfair to attribute any difference in mean outcomes to the treatment because we wouldn't be able to separate out what part of this difference was due to the program and what part of the difference was due to the baseline differences between groups.

### 20.2.2 Potential outcomes and causal quantities

We can conceptualize the assumptions required for causal inference as a formalization of this idea of fair comparisons. Critical building blocks in this formalization are *potential outcomes*. In the cholesterol example, the potential outcome $Y_{0i}$ represents the cholesterol level we would see for individual $i$ if they *did not* participate in the program. $Y_{1i}$ represents the cholesterol level we would see for individual $i$ if they *did* participate in the program. Potential outcomes allow us to formally define a causal effect for person $i$ as the a comparison between them, $\tau_i = Y_{1i} - Y_{0i}$. This effect represents how different the cholesterol level would be for individual $i$ if they *had* participated in the program compared to if they *had not*.

In practice, however, we can never measure both potential outcomes at the same point in time for the same person because we cannot observe that person both in the world where they received a treatment and the world where they did not. Therefore, if we denote the binary treatment received by individual $i$ as $Z_i$, we can express the observed outcome, $Y_i$, as a function of the potential outcomes, $Y_i = Z_i Y_{1i} + (1 - Z_i)Y_{0i}$, such that $Y_{1i}$ is revealed for those who receive the treatment and $Y_{0i}$ is revealed for those who do not. We can perhaps imagine a situation where we could clone individual $i$ to create individual $j$ just at the moment of exposure to the treatment and have version $i$ take the treatment and version $j$ refrain. This would create a fair comparison when we compared their outcomes down the line because we would be assuming that both the individual and their clone would have had the same $Y_0$ and $Y_1$. Specifically we could assume $Y_{0i} = Y_{0j}$ and $Y_{1i} = Y_{1j}$. Therefore to estimate the causal effect for individual $i$, $\tau_i = Y_{1i} - Y_{0i}$, and even though their $Y_{0i}$ would be missing, we could just substitute $Y_{0j}$ which *is* observed (since individual $j$ did not receive the treatment).

Most causal inference procedures try to mimic this situation but generally aim to estimate *average* treatment effects such as the mean of individual causal effects over a sample or population. We can express such an average effect generically as $\mathrm{E}[Y_{1i} - Y_{0i}]$. Some researchers may be interested instead in the average treatment effect for the type of individual who we observed to receive the treatment (or participate in the program). This quantity is referred to as the average effect of the treatment on the treated (ATT) and is formalized as $\mathrm{E}[Y_{1i} - Y_{0i} \mid Z_i = 1]$. This quantity may be of particular interest if we don't expect that our full control group would be eligible for or interested in the treatment. A reciprocal version of the ATT is the effect of the treatment on the controls, ATC, $\mathrm{E}[Y_{1i} - Y_{0i} \mid Z_i = 0]$, which captures the effect of the treatment for those we don't observe to take the treatment (or participate in the program). This may be of particular interest in situations where policy makers or practitioners would like to expand eligibility or incentivize different types of people to participate in a program or receive a treatment.

### 20.2.3 Assumptions

If our goal is to estimate the average causal effect for a group of individuals, we need to create fair comparisons for the group. What would be the most pristine way to accomplish this? Suspending disbelief for a moment, let's suppose we could clone everyone who was willing to participate in the study. Each of the original participants could be exposed to the treatment but none of the clones would be exposed. What are the implications for the assumptions needed for causal inference?

#### 20.2.3.1 All confounders measured

In this hypothetical study with clones, one nice feature is that the everything about the original sample and the clones would be the same except for their treatment (and everything subsequent to the treatment). Since potential outcomes exist conceptually even prior to the treatment being administered, this implies that the distribution of potential outcomes would be the same across the treatment and the control groups,

$$p(Y_0, Y_1 \mid Z = 0) = p(Y_0, Y_1 \mid Z = 1).$$

This property is more commonly expressed as an independence statement

$$Y_0, Y_1 \perp Z.$$

While it is not possible to actually implement our hypothetical clone design, a completely randomized experiment does a good job of at least replicating this independence property because the randomization eliminates any *systematic* differences between the treatment and control groups. Unlike with our clones, the randomized experiment is still most useful for estimating average treatment effects. However, this does not guarantee that it will yield accurate *individual-level* treatment effect estimates.

Of course with small sample sizes a randomized experiment may still result in groups that differ simply by chance. While results will still be technically unbiased, any given treatment effect estimate may still be far from the truth[1]. Randomized block experiments, in which the randomization occurs within strata or blocks defined by covariates, can help to address this by ensuring perfect balance with respect to the blocking variables. Generally speaking, in randomized blocks experiments, independence is only achieved within blocks therefore these experiments satisfy a more specific assumption,

$$Y_0, Y_1 \perp Z \mid W,$$

where $W$ denotes the blocks. For example, in a diet and exercise intervention we might seek to randomize individuals after grouping them based on their starting cholesterol levels.

Of course many questions can't be be addressed by randomized experiments due to any combination of logistical, financial, and ethical reasons. In that case researchers hope that the covariates they've measured essentially act like blocks in a randomized block experiment. That is, they hope that observations that have similar values on all their covariates have potential outcomes that are similar as well, regardless of their treatment assignment.

This assumption can be expressed formally as

$$Y_0, Y_1 \perp Z \mid X,$$

where $X$ now denotes pre-treatment covariates. The intuition here is that for two groups (treatment and control) with the same values on all the pre-treatment variables, $X$, we assume they have, in effect, been randomized to the groups. This is basically the same assumption as is invoked by the randomized block experiment. The difference is that in a randomized block experiment the $W$ are known and the assumption should hold in a pristine implementation of the design. In on observational study, on the other hand, researchers must make a leap of faith that their pre-treatment covariates, $X$, are sufficient to achieve this conditional independence. If $X$ represents all confounders (informally, variables that predict both treatment and outcome), this assumption should be satisfied.

This assumption is referred to by many different names depending on the discipline and subfield. These include "ignorability," "selection on observables," "all confounders measured," "exchangeability," the "conditional independence assumption," and "no hidden bias" [5–8, 23]. In this chapter we will refer to this as the "all-confounders-measured" assumption.

Due to the critical role of this assumption in estimating unbiased treatment effects, the first step in many causal inference approaches is to try to ensure sure that is satisfied by including all *potential confounders* in $X$.[2] The second step, which is the focus of many of chapters in this book, is to figure out how to condition on these variables without making excessive additional assumptions. We'll illustrate some of the complications involved in this step in the next section with a hypothetical example, but first we discuss a related assumption.

#### 20.2.3.2 Overlap

Another property of our hypothetical example with clones is that for each individual in our dataset there would be someone else with the exact same values of all pretreatment covariates (including confounders) but who received a different treatment. We might relax this to say that this neighbor

[1] For more discussion see [4].

[2] The idea of including all *possible* confounders is in tension with the desire to avoid "overfitting", discussed in Section 20.3.1: it cannot be fully known whether a variable is truly a confounder or a randomly correlated covariate, so that including additional but unrelated pre-treatment covariates may reduce generalizability. Another concern in including many potential confounders in a model is "bias amplification," which can occur when some true confounders are missing [10–12] and covariates are included that are strongly predictive of the treatment but not the outcome. However, this may be rare in real-world data, such that it is generally preferable to condition on most pre-treatment covariates [13]. Our advice would be to always include every pre-treatment covariate that is believed to be predictive of the outcome, particularly if it is also related to the treatment. If overfitting remains a concern use regularized/Bayesian models or perform a variable selection step.

would have to be in a sufficiently close neighborhood of the covariate space. Thus, by design, each individual would have an "empirical counterfactual" in the dataset. A more general formalization of this property is that, for every $X$ and for $z \in \{0, 1\}$, $0 < \Pr(Z = z \mid X) < 1$. Conceptually we can think of this expression as requiring that in every neighborhood of the covariate ($X$) space spanned by our sample there has to be a positive (non-zero) probability of having both treated and control units.

In theory a completely randomized experiment should create overlap by design since the multivariate distributions of all pre-treatment variables (measured of not) should be the same across groups. Of course, in practice, if one were checking, it would be difficult to achieve perfect overlap even for a moderate number of variables simply due to sparsity [14]. However it turns out the overlap assumption is technically stronger than what we need to perform inference. A more precise requirement is that we have overlap with respect to our *true confounders*, $X^C$, as in $0 < \Pr(Z = z \mid X^C) < 1$. We'll return to this idea later in the chapter because BART affords some advantages with regard to this goal [15]. Since a completely randomized experiment has no confounders the requirement is satisfied trivially. A randomized block experiment would need to satisfy the overlap assumption with respect to its blocks.[3] Again this should be satisfied by design since randomized block experiments require a non-zero probability of assignment for each treatment variable in each block.

In an observational study, however, there is nothing to guarantee that overlap exists across treatment groups with respect to confounders. There may be certain type of people who will never participate in a program or will always be exposed to a treatment. If the overlap assumption does not hold, there may be some observations on our data set for which we simply don't have enough information about their counterfactual state to try to make inferences about them.

To push this to the extreme imagine what would happen in a study if all confounders are measured but overlap is violated. For example, suppose that we have a study where individuals are assigned to receive a treatment based on their age. Specifically, suppose that all individuals over age 50 receive the treatment and all individuals 50 and younger do not. Further suppose that within each group, treatment assignment has no impact on the potential outcomes; if the age restriction on treatment did not exist, the experimental design would be valid. Thus, even though all confounders are measured there is no overlap in the age distribution across treatment and control groups. In this situation arguably none of the observations would have empirical counterfactuals. If you wanted to understand the effect of the treatment for the individuals less than 50 you would be hampered by the absence of treated units in this age range to provide data on the missing $y_1$'s. If you wanted to understand the effect of the treatment for individuals older than 50 you would be hampered by the absence of control units in this age range to provide data on the missing $y_0$'s. Without further assumptions your only hope would be to focus on those observations closest to the threshold.[4]

#### 20.2.3.3 SUTVA

While we won't devote much time to it in this chapter, one of the most important assumptions in causal inference is the Stable Unit Treatment Value Assumption (SUTVA) [18]. The basic idea is that we need to assume that each person's potential outcome is a function solely of their own treatment assignment, not the treatment assignment of anyone else. This portion of SUTVA is sometimes referred to as the "no-interference" assumption. This assumption also encapsulates the idea of "consistency." This can be formalized as the idea that $Y_{aj} = Y_j$ when $A_j = a$ [19]. This reflects the idea that if the observed value of $Y$ for an individual $j$ who received treatment equal what it would have been if the treatment was "set to" $a$. In other words, we assume that the manner in which treatment was set to $a$ is irrelevant. This is sometimes referred to as the "no-multiple-versions-of-treatment" assumption [20].

[3] Not all blocks act as confounders, however, so this is actually also too strong a statement.

[4] For clever ideas of how you *can* estimate a causal effect in this situation, you can read about regression discontinuity designs [4, 16, 17].

## 20.3 Regression for Causal Inference

We illustrate the use of regression for causal inference by introducing a hypothetical example that we will augment to illustrate key features of the approach. This example imagines that a few years ago a large sporting equipment company released a new marathon running shoe called the hyperShoe with the claim that runners wearing the hyperShoe would be able to run faster without changing their fitness or running technique. In effect, the sporting company made the claim that using the hyperShoe caused reductions in marathon race times. Upon release, the hyperShoe was marketed to specific subset of runners. Thus, runners using the high performance shoes likely differ with respect to characteristics that also would be expected to predict their running performance. For instance, those who purchase the shoe are likely to be a different age and may be more serious about running. While the company claims that the high performance running shoes result in faster race times, skeptics argue that self-selection may explain the differences in performance. In other words if higher-performance runners are more likely to buy the shoes, then any differences in running times might simply be a result of the differences in the types of runners who tend to wear the shoes. In technical terms potential confounders need to be ruled out before reaching this causal conclusion.

To better understand the role of confounders, let's focus first on age, since it is easy to measure and is one of the strongest predictors of running time. Suspending disbelief for a moment, if we assumed that age were the *only* confounder when estimating the causal effect of the hyperShoe on running times, what would it mean to make fair comparisons? In this hypothetical, we would need to compare individuals of the same age but who differ with regard to whether they wore the hyperShoe. How does regression allow us to make this comparison?

Regression provides one way to condition on confounders in an attempt to create fair comparisons. Most students who take introductory statistics courses are taught that linear regression should not be used to draw causal conclusions. After all, they are told, and rightly so, "correlation is not causation." But that doesn't really get to the heart of the matter since alternatives like matching and weighting also just estimate (conditional) correlations or associations and also require additional assumptions to identify causal effects. It is critical to understand under what conditions a regression could recover a causal effect.

Let's start with a version of our hypothetical example in which age is linearly related to running time. Figure 20.1 displays what this might look like. Here lighter dots represent "treated" units (those with the hyperShoe) and darker squares represent control observations. The lighter line represents the relationship between $Y_1$ and age. The darker line represents the relationship between $Y_0$ and age. These lines are sometimes referred to as the "response surface". The vertical difference between the two lines represents, $\tau$, the treatment effect at each level of age. Since the treatment effect is constant across age (reflected in the parallel lines), the expected treatment effect for any given person is the same as the average across the sample. The following regression equation, displayed by the dashed lines in the plot, provides a good fit to the data.

$$Y = 190.28 + .49X - 10.21Z + \epsilon,$$

where $Y$ denotes running time, $Z$ denotes use of the hyperShoe, and $X$ denotes age. But is the coefficient on $Z$ an estimate of the causal effect of the hyperShoe on running time?

Recall that in this example we are assuming that age is our only confounder. Thus the assumption of "all confounders measured" is satisfied. If, additionally, our parametric model is correct, we can say that

$$E[Y_0 \mid X] = E[Y_0 \mid X, Z = 0] = E[Y \mid X, Z = 0] = \beta_0 + \beta_1 X$$

and

$$E[Y_1 \mid X] = E[Y_1 \mid X, Z = 1] = E[Y \mid X, Z = 1] = \beta_0 + \beta_1 X + \tau.$$

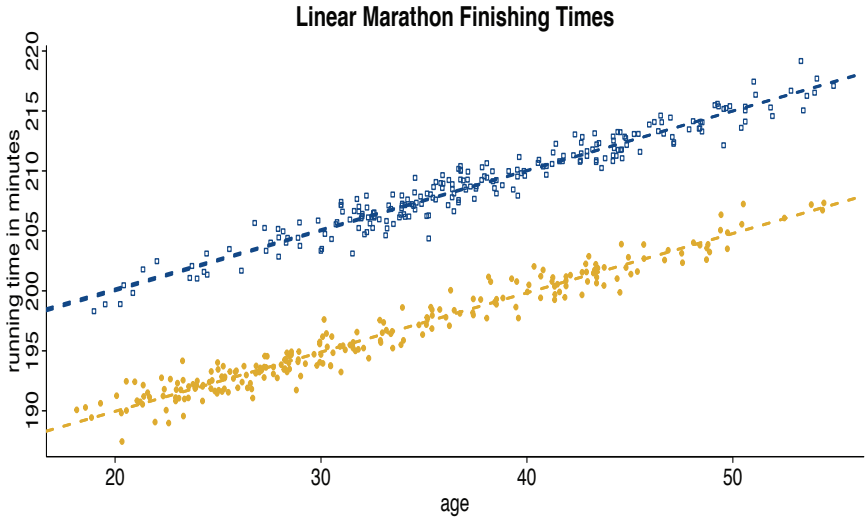


**FIGURE 20.1**
Hypothetical data on age and running times for a marathon. The solid line displays the true relationship between age and running time which, implausibly, is linear in this example. Lighter dots represent "treated" units – those with the hyperShoe – and darker squares represent control observations. The lighter line represents the relationship between $Y_1$ and age. The darker line represents the relationship between $Y_0$ and age.

That is, in this situation our regression model is also a causal model. Thus $\hat{\tau} = -10.21$ should be an unbiased estimate of the true treatment effect. In this case since we simulated the data we know that the true treatment effect for this example is $-10$. Interpreted causally, the hyperShoe decreases race time by 10 minutes.

In sum, if age is the only confounder *and* if the linear model is appropriate, then a linear regression model *can* recover the causal effect of the hyperShoe on running times. Unfortunately, neither of these assumptions is generally appropriate. We focus for the time being on the modeling component, and return to the foundational assumptions for causal inference in a later section.

### 20.3.1 Regression trees vs. linear regression

Unfortunately the real-world relationship between age and running time is not accurately represented in the hypothetical data presented in Figure 20.1 above. Figure 20.2 plots hypothetical age and marathon time data that better reflects the relationship documented in the literature [21, 22] as a blue line. The darker line, in contrast, represents what this relationship might look like if the hyperShoe was successful in reducing running times. Unlike in our previous example, a linear regression fit to the data (displayed with the dashed lines) fails to correctly represent the true relationship between age and running time for each potential outcome. The linear regression estimate of the causal effect is about $-3.66$, with a $95\%$ confidence interval of $(-2.98, -4.34)$. However, the true ATT, ATE, and ATC for this sample are $-3.85$, $-4.42$, and $-5.00$, respectively. The linear regression confidence interval, which is best suited to estimate the ATE, captures the ATT but not the ATE or ATC. What's going on?

One obvious problem is that the true response surface is not linear! So we wouldn't expect a linear regression to represent it well. But there is another problem that exacerbates the issue. Not all parts of the of the "covariate space" (here, the range of the age) are well represented by both treated and control observations. Consequently, the model underperforms most severely in regions of the

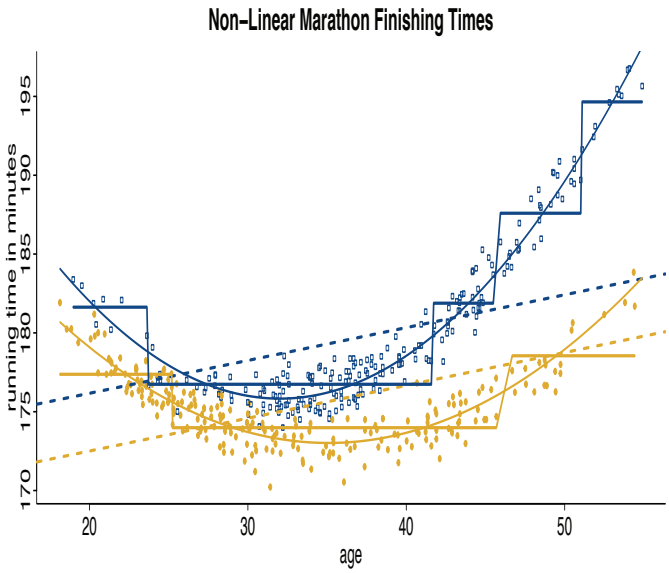


**FIGURE 20.2**
Hypothetical data on age and marathon running times in minutes. The solid curves (lighter for $Y_1$ and darker for $Y_0$) display the true relationship between the confounder and the potential outcomes. The dashed lines (with the same color mapping) represent a linear regression fit. The long-dash lines (with the same color mapping) represent the fit from a regression tree. Lighter dots are treated observations. Darker squares are control observations.

covariate space where counterfactual units are rare. While the treatment group is mostly strongly represented in the age range from 18 to 50, the control group more closely spans the full age range from about 20 to 55. Thus there are more empirical counterfactuals for the full treatment group than vice-versa.

Is there a simple way of providing a better fit to this response surface? Recall that regression is just a way to summarize information about how average outcomes of the response variable vary across subgroups defined by the covariates in our sample. In our current example we want to understand how marathon running times vary with subgroups defined by age. Linear regression does this in a way that places strong constraints on how these means are related to each other. Can we fit this relationship using a regression model that makes fewer assumptions?

A regression tree fit to these data would deconstruct the problem a bit differently than linear regression. Regression trees form subgroups within the dataset such that the within-subgroup variance in the outcome variable across subgroups is minimized (see chapter 9 of The Elements of Statistical Learning (ESL) by Hastie, Tibshirani, and Friedman (2009) [23]). The first step is to split the dataset into two subgroups. In a regression tree fit to our example data, the first split divides those individuals who are younger than 46 from those individuals older than 46. If we allow for further splits, the tree will continue to subdivide individuals by age and by who wore and did not wear the hyperShoe. This splitting process is repeated until a stopping condition is met, in our case until eight subgroups, or terminal nodes, are found. The tree with this stopping rule is displayed in the right panel of Figure 20.3 which shows the decision rules for each split and the mean in each terminal node. For instance the right most terminal node shows a mean of 195 which is the average outcome for individuals in our sample who did not receive the treatment (wear the shoe) and who are age 51 or above. The fit from this regression tree is summarized by the mean outcomes for the terminal node

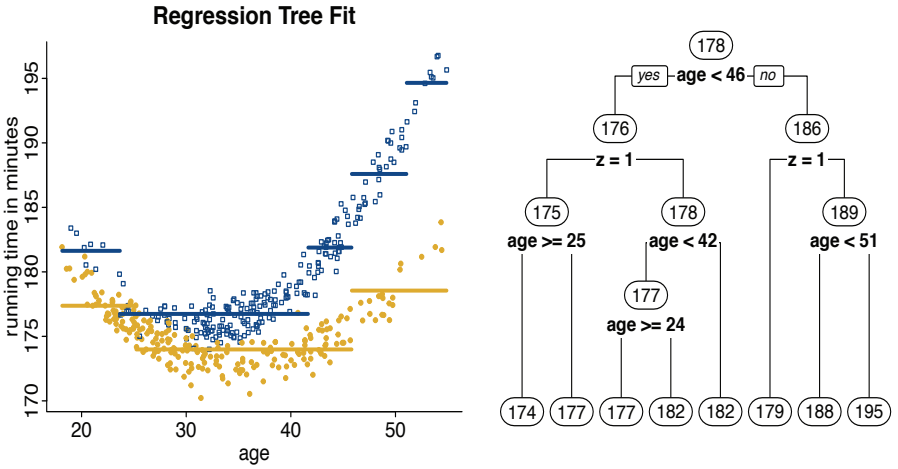


**FIGURE 20.3**
The left side of this figure displays the regression tree fit to the response surface which is represented by eight subgroup means. This means are displayed as horizontal line segments spanning the ages of the individuals in their subgroups and with darker colors for control observations and lighter for treated observations. The right side displays the branching and terminal nodes from the corresponding regression tree. In this tree "z" represents the binary treatment variable.

subgroups. These means are visualized by the horizontal lines (step function) in Figure 20.3. Notice how this model is better able to follow the curve of the response surface displayed on the left.

The flexibility of the regression tree fit is appealing. What are the downsides? We didn't mention above specifically how we decided to stop "growing the tree" (i.e. creating more subgroups). This is a tricky decision. If we stop too early we risk creating too crude a fit to the data. This is apparent in the fit in Figure 20.3, where it is often further from the observed data than we would like. On the other extreme we could allow the tree to grow so large that the fit yields a different "mean" for every observation. This model fit would be terrific at predicting the outcome within the current sample but is not likely to predict well at all in a new sample. This phenomenon is referred to as "overfitting." For a discussion of overfitting, see Chapter 7 of ESL.

In practice this tension is resolved by adjusting tuning parameters (also called hyperparameters in some fields) for the algorithm that govern features such as the number of observations required for a terminal node to be allowed to split, the minimum deviation within a terminal node to prevent further split, and the maximum depth for a tree. Typically these parameters are chosen via cross-validation (see chapters 7 and 9 of ESL). However cross-validation can only be directly used to understand to guard against overfit with regard to our observed data. It cannot be used to understand potential overfit with regard to unobserved counterfactuals.

There are several other downsides to regression trees. First, when we have multiple covariates, they aren't able to effectively capture additive effects well and tend to overemphasize multi-way interactions. Second, they don't directly estimate our uncertainty about our predictions or fit. This latter issue can be addressed using bootstrapping but that comes at a high computational cost. Finally, they tend to have a high variance – slightly different datasets can yield dramatically different trees.

### 20.3.2 Boosted regression trees

Boosted Regression Trees (chapter 10 of ESL) emerged as a way to address these issues of overfitting, difficulty in capturing additive structure, and overemphasis on high-level interactions. The idea

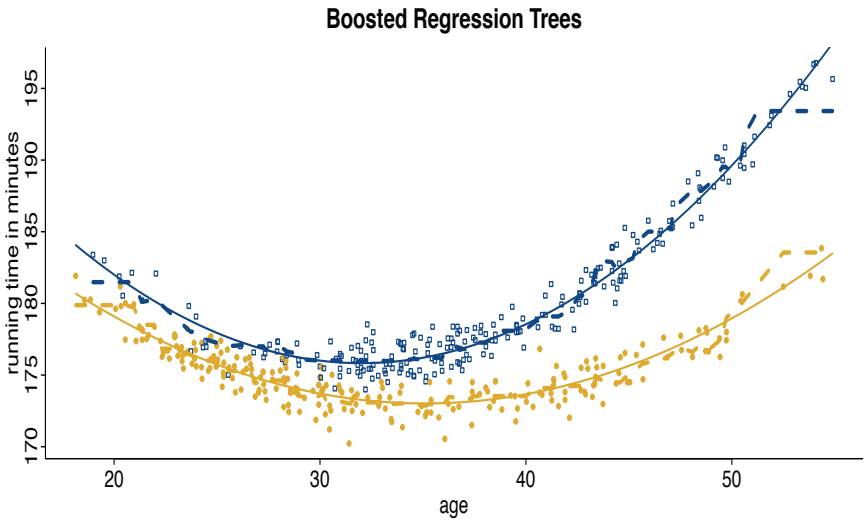


**FIGURE 20.4**
This figure displays the same response surface as in the previous figure (light for treated and Y(1) and dark for controls and Y(0)) but now with a boosted regression tree fit displayed as dashed lines.

is very clever! Instead of representing the fit to the data through a single regression tree, boosted regression trees represent the fit to the data through the sum of fits from multiple small regression trees. How does this work?

Consider our example above. A single tree fit would form subgroups based on treatment assignment and age as displayed in Figure 20.2. The residuals from that fit represent the variation in the outcome that has yet to be explained (which is a lot at this point!). What if we fit another tree to those residuals to try to explain more? Now new residuals can be formed by subtracting the fit from the second tree from the residuals from the previous step. This process can be repeated many times and at each step more of the unexplained variation in the outcome can be explained. Using this strategy, model complexity is controlled by using only small trees at each step, also known as "weak learners," tree predictions are combined using a weighted average or sum, reducing variance and forming an "ensemble," and by limiting the overall number of trees overfitting can be avoided.

Figure 20.4 displays a fit from a boosted regression tree with 100 trees. There is marked improvement to the fit of the response surface relative to the smaller tree in 20.3. However, note that the tree fit starts to depart from the true response surface when the treatment or control observations become relatively scarce. Traditional, machine-learning-style fits typically provide no way of being alerted to the increased uncertainty in this part of the covariate space.

While boosted regression trees were an important step forward as compared to standard regression trees, they still require choice of tuning parameters (additionally now including the number of trees included in the sum of trees) and fail to resolve the issues regarding uncertainty quantification.

## 20.4 Bayesian Additive Regression Trees

Bayesian Additive Regression Trees (BART) is a model that addresses some of the outstanding issues with boosted regression trees. It handles overfitting in a more principled, flexible, and data-driven way and also provides coherent uncertainty intervals.

The mean structure of BART is the same as the boosted regression tree. However, this structure is then embedded in a likelihood framework so that the fit of the model is considered to be a random variable and hence has a distribution. Given the importance of the treatment in this context we will write the BART for causal inference model explicitly incorporating the treatment variable in the notation. In addition, an error term is added to reflect deviations between the fit from the model and our observed values of the outcome (which reflects our uncertainty). We can formalize the model as

$$Y_i = f(\mathbf{x}_i, z) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \tag{20.1}$$

$$f(\mathbf{x}, z) = g(\mathbf{x}, z, T_1, M_1) + g(\mathbf{x}, z, T_2, M_2) + \cdots + g(\mathbf{x}, z, T_m, M_m). \tag{20.2}$$

In this equation each of the $g$ functions represents the fit from an individual tree, $T_h$ represents the structure of the $h^{\text{th}}$ trees, and the corresponding $M_h = (\mu_{h1}, \mu_{h2}, \ldots \mu_{hb_h})'$ represents the set of subgroup means corresponding to the $b_h$ terminal nodes of tree $h$.

### 20.4.1 BART prior

The key innovation in the model rests in the prior. Three elements of the BART model require prior specifications: the trees, the subgroup means, and the error term. While technical details can be found elsewhere (see BART: Bayesian Additive Regression Trees by Chipman, George, and McCulloch (CGM, 2010; [2]), we describe here the intuition behind and implications of the prior distributions.

#### 20.4.1.1 Prior on the trees

The goal of the prior specification on the trees is to keep each of the individual trees relatively small. In fact, the default prior specification results in a prior probability for trees with only a single, terminal node of $0.05$. Similarly, the prior probabilities for trees of sizes 2, 3, 4, and 5 or more terminal nodes are .55, .28, .09, .03, respectively. By expressing a preference for small trees, this prior avoids over-reliance on any single tree and also discourages high-level interactions. However, if the data suggest that larger trees are warranted the model will accommodate this. For instance in our example data a BART model constrained to have only a single tree in the response model has a median depth across iterations of 6. In a setting with more variables and a more complex structure, however, we might expect the average tree size to be larger.

#### 20.4.1.2 Prior on the means

The prior on the means is also designed to help avoid overfitting. This is achieved by specifying a reasonable prior for the overall fit and then "reverse engineering" the implications for the means from individual trees. What is a reasonable prior for the overall fit? CGM achieve this by standardizing the outcome data so that it all lies between $-.5$ and $.5$. Given this, a reasonable prior for the expected response for the fit would assign high probability that the sum of all of the trees lies between these two numbers. The default prior sets this probability to $95\%$.

Remember, however, that the total expected value is the sum of $m$ means from individual trees. The prior over these means can be derived from the overall prior. Suppose that we express the sum of the individual trees as $\text{N}(0, m\sigma_t^2)$, where $\sigma_t^2$ is the variance of the prior mean for any given tree. For this sum, by default, to have $95\%$ probability of being within $-.5$ and $.5$ while treating the prior contribution of each tree as independent and identically distributed, $\sigma_t$ has to be equal to $.5/2\sqrt{m}$, where $2 \approx 1.96 \approx \Phi^{-1}(0.975)$. Consequently, $\sigma_t$ is a hyperparameter which determines the sensitivity of the nodes to the data. The default values work very well for continuous response variables; however, $\sigma_t$ may require adjustment, crossvalidation, or the imposition of a hyperprior to avoid overfitting.

#### 20.4.1.3 Prior on the error term

The prior on the error term should reflect the level of uncertainty we expect will remain after fitting the model. Rather than specifying an informative prior[5], CGM calibrate the prior using the residual error from a linear regression fit as a benchmark. Specifically the $\chi^2$ prior assumes that there is $90\%$ chance that the BART residual standard deviation will be less than that estimated by a standard linear regression.

Why does this make sense? Well suppose that the true model is linear. Then we would expect BART to have a similar residual standard deviation, though possibly slightly larger since it has to approximate a straight line with step functions. However, if the true response surface is nonlinear, then we would expect BART to have a smaller standard deviation due to the closer fit. In essence this prior assumes that in $90\%$ of the examples where we will fit BART standard linear regression model will not provide a better fit, as defined by the residual standard deviation.

### 20.4.2 Gibbs sampler for BART

As a Bayesian algorithm, BART is fit by combining a statistical likelihood with priors on parameters to define a posterior distribution. Posteriors that do not have simple, named distributions are often summarized by sampling procedures. These samples can be drawn by defining a random walk whose stationary distribution is that of the target, that is by using a Markov chain Monte Carlo (MCMC) algorithm [24]. The algorithm starts with parameters initialized in some fashion – often drawn from their prior distribution or given reasonable, dispersed starting points – and proceeds by iteratively updating them from their current state to new ones in a stochastic fashion. The initial portion of such walks are often called "warm-up" and are not used in inference. During warm-up, the sampler progresses from the starting point to, ideally, the stationary distribution. Multiple, independent chains are often run as a diagnostic: when the random walks from many chains are all found to be in the same area of the posterior, the sampler is said to have "converged" to the stationary distribution.

In particular, BART is implemented as a Gibbs sampler, where the random walk proceeds by sequentially sampling from the conditional posteriors of individual parameters, given all of the others [25]. Since the likelihood for BART is Gaussian and the contribution of each tree is conditionally independent of all of the others, a Gibbs sampler for BART gives rise to an algorithm called "Bayesian backfitting": each tree can be randomly manipulated while fit against the residual of the response and every other tree. For details, see CGM. Finally, as a Gibbs sampler with a Gaussian likelihood, BART can be embedded in hierarchical models with components conditionally sampled, given different parameters. For example, see `stan4bart` [27], which fits multilevel models with both BART and so-called random effect mean structure. We discuss this extension in more detail below.

## 20.5 BART for Causal Inference

Just as a linear regression or regression trees could be used to estimate a causal effect, BART can be used as well. This section describes the steps involved in a basic implementation and available software.

[5] We avoid using the term "uninformative," as all priors incorporate some choice on where to distribute probability mass or density. For example, a so-called "flat" prior on a standard deviation expresses a strong preference for small valued variances.

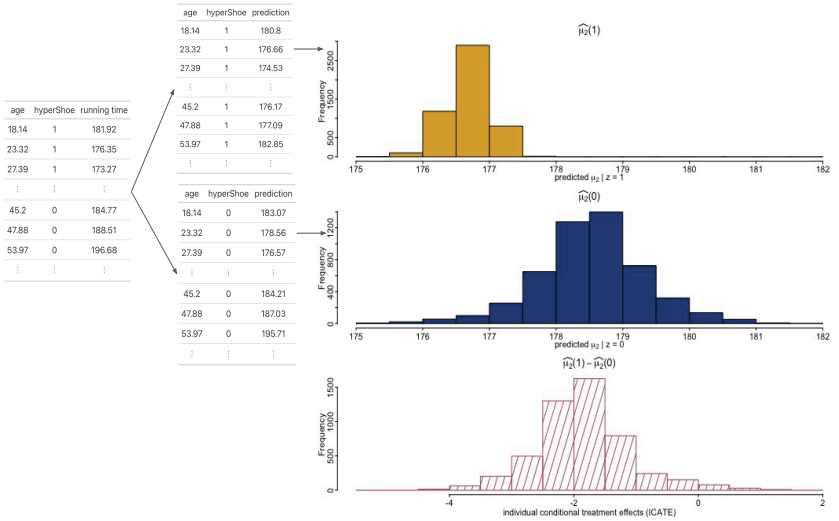


**FIGURE 20.5**
This figure displays how BART is used for causal inference. BART is fit to the original dataset (on the left). Predictions are made for two altered versions of that dataset (in the middle): one where all observations are assignment the treatment (top) and one where all observations are assigned to the control (bottom). These represent predictions of $Y_0$ and $Y_1$ for each person (last column of middle datasets). Posterior predictive intervals for each potential outcome for the 2nd individual in the dataset are displayed on the far right. The top plot is a histogram showing the empirical posterior distribution of $Y(1)$ for that individual. The darker histogram below it shows the empirical posterior distribution of $Y(0)$ for that individual. The difference between these distributions is the posterior distribution of the treatment effect for the 2nd individual (bottom-most histogram). These individual-level treatment effect distributions can be combined to estimate an average treatment effect for any of a variety of average treatment effects.

### 20.5.1 Basic implementation

A simple BART approach to causal inference implicitly involves several steps that are automated in current software (`bartCause`, described below). First, the algorithm is fit to the data using the model and fitting procedure described above. Second, predictions for the outcome are made for each observation in the original dataset, as well as a new "counterfactual" dataset. Intuitively we can think of that algorithm as imputing the missing potential outcomes in the dataset. In this way we can obtain predictions for each individual both for their observed treatment assignment and their counterfactual assignment.[6]

Since BART is a Bayesian algorithm, it provides not only a prediction of $Y_0$ and $Y_1$ for each person, but a full posterior predictive distribution for each of these potential outcomes. These distributions can be combined for a given individual, $i$, by differencing the two distributions to obtain a posterior distribution for $Y_{1i} - Y_{0i}$, individual $i$'s causal effect. Figure 20.5 illustrates these steps and displays posterior distributions for the second observation in the dataset of our hypothetical example. Critically, posterior distributions for individual level causal effects can then be averaged to obtain the posterior distribution of the causal effect for the full sample or any subset thereof.[7]

[6] We describe it this way to make the intuition clear. In reality the computations only require that we fit on one dataset and make make predictions for that dataset and a version where the treatment assignment for all observations is flipped.

[7] Since these distributions condition on the covariates in our analysis sample, the average causal effect technically represents a conditional average causal effect rather than the sample average causal effect. Sample average effects can be obtained by

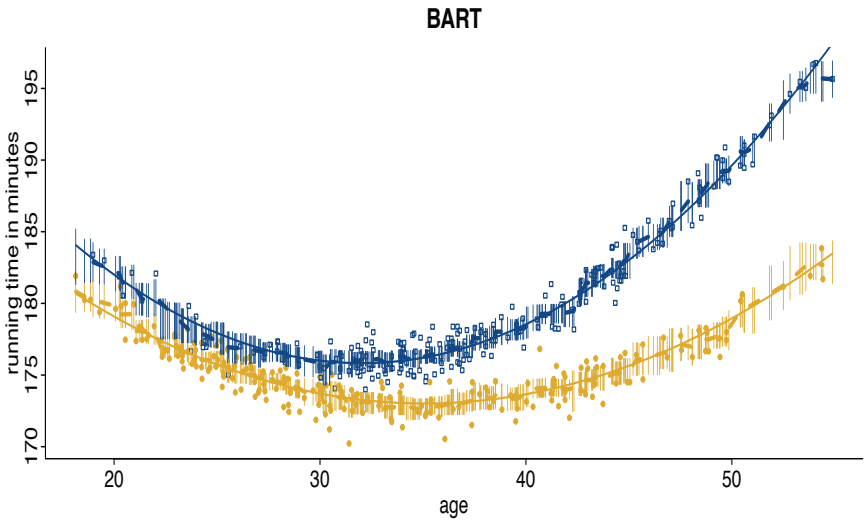


**FIGURE 20.6**
This figure displays the same response surface as in the previous figure (darker curve for treated and lighter curve for control) but now with a BART fit overlaid as dashed lines. 95% posterior uncertainty intervals (vertical lines) for all individuals (again with lighter lines for the treated and darker lines for the controls) are calculated by using normal approximations to their empirical marginal posteriors.

We present results of the BART fit to the data in our hypothetical example in Figure 20.6. This plot reproduces the true response surface and observations from Figure 20.4 above but instead overlays a BART fit. There are several ways to display this fit but we choose here to display two uncertainty intervals for each observation, $i$ on the plot: (a) a 95% uncertainty interval for $E[Y_{0i} \mid X_i]$ and (b) a 95% uncertainty interval for $E[Y_{1i} \mid X_i]$.

### 20.5.2 Software: `bartCause`

A variety of packages currently exist in the R programming language to implement BART ( [27, 29–35]), encompassing a multiplicity of applications and algorithmic adaptations. The R package bartCause allows for estimation of treatment effects using BART as described above and additionally accommodates the extensions discussed in the following section. It accepts as input the treatment and response variables as well as a list of confounders. It then creates a suitable counterfactual dataset to provide predictions of potential outcomes and their uncertainty.

The main function of the bartCause package is `bartc`. `bartc` is designed to be relatively similar to other model fitting functions in R, like `lm`, with the exception that it can fit *two* models instead of one. Consequently, it breaks the typical `formula` argument into three parts: `response` for the name of outcome variable, `treatment` for the name of the treatment, and `confounders` for the names of the covariates included to try to satisfy ignorability. The treatment and confounders are used to predict the outcome as described above. The confounders are also used in a model to predict the treatment as a way to flexibly estimate propensity scores. The propensity scores can be useful as an additional predictor in our outcome model or in calculating weights to help balance treatment and control groups.

---

using the observed factual outcome, and draws from the posterior predictive counterfactual distribution [80]. The uncertainty of population average effects can be obtained by using the posterior predictive distributions for both potential outcomes.

Additionally, bartCause accepts the standard model function fitting argument of a `data` object where it will look to resolve symbols, which can be a data frame or list with named elements. In other words, it takes inputs that are similar to the `lm()` and `glm()` functions in R. Since BART will look for a flexible, non-parametric relationship between the confounders and the treatment or response, the `confounders` themselves need merely be named. Deriving from the R model fitting syntax, they can be separated with a "+" sign, however, this is interpreted figuratively as "include this variable" and does not indicate a linear relationship. The `estimand` argument is also important to specify up front because the methodology differs for some approaches when targeting the `"ate"`, `"att"`, or `"atc."`

In the context of our illustrative example we could use the following command:

```
bartc_fit <- bartc(running_time, hyperShoe, age,
                   data = dat, estimand = "att",
                   seed = 0)
```

Once a model has been fit, callisng `summary` on it yields inferential results, including by default an estimate of the relevant population average treatment effect (in this case the population average effect of the treatment on the treated, PATT):

```
> summary(bartc_fit)
Call: bartCause::bartc(response = running_time, treatment = hyperShoe,
            confounders = age,  data = dat, estimand = "att",
            seed = 0)

Causal inference model fit by:
  model.rsp: bart
  model.trt: bart

Treatment effect (population average):
    estimate     sd ci.lower ci.upper
att   -3.866 0.1616   -4.182   -3.549
Estimates fit from 464 total observations
95% credible interval calculated by: normal approximation
  population TE approximated by: posterior predictive distribution
Result based on 500 posterior samples times 10 chains
```

In addition, bartCause includes a number of convenience plotting functions, all of which take a fitted model as their first argument.

- `plot_sigma`: for continuous responses only, produces by-chain trace plots of the residual standard deviation, where the $x$ axis is the sample number and the $y$ axis is the value; for use in assessing model convergence
- `plot_est`: by-chain trace plots of the estimand
- `plot_indiv`: histograms of individual-level quantities, including treatment effects and posterior means
- `plot_support`: scatter plots of individual-level quantities, with observations highlighted by the evidence of their common support, as discussed below

Finally, advanced users can access the posterior samples directly, using the `extract`, `fitted`, and `predict` generic functions. These can be useful for obtaining subgroup estimates, weighted averages, or for conducting additional diagnostics. bartCause can also be accessed now in a more user-friendly software package, thinkCausal, that additionally incorporates educational components to help the user understand the foundational concepts involved (https://apsta.shinyapps.io/thinkCausal/).

## 20.6 BART Extensions and Other Considerations for Causal Inference

The discussion above focused on average treatment effects for independent, identically distribution data where we assume that the all-confounders-measured and the overlap assumptions hold. What happens if these conditions aren't met or we want to estimate more complex treatment effects? This section briefly explores these issues.

### 20.6.1 Overlap, revisited

We revisit the overlap assumption to better the understand its implications vis-a-vis the BART approach to causal inference and other common causal inference strategies. To accurately estimate treatment effects, this identification strategy relies on its ability to accurately predict counterfactual values for the inferential observations (those represented in the estimand of interest). For instance, suppose the focus is on the effect of the treatment on the treated, $\mathrm{E}[Y_1 - Y_0 \mid Z = 1] = \mathrm{E}[Y_1 \mid Z = 1] - \mathrm{E}[Y_0 \mid Z = 1]$. We know a lot about $\mathrm{E}[Y_1 \mid Z = 1]$ because we observe $Y_1$ for all the members of the treatment group. However, we have no direct measures of $\mathrm{E}[Y_0 \mid Z = 1]$ because $Y_0$ is not observed for anyone in the treatment group. However, if all-confounders-measured holds, we can capitalize on the following equivalency: $\mathrm{E}[Y_0 \mid Z = 1, X = x] = \mathrm{E}[Y_0 \mid Z = 0, X = x]$. At a conceptual level, for each treatment group member, we use the information from control group observations that have the same values of $X$ as the treatment group member to predict their $Y_0$. In practice, since it is unlikely that there are observations with exactly the same values on all variables, most matching methods matched on a lower-dimensional representation of the covariates, such as the propensity score. BART instead makes a prediction using its extremely flexible model.

In practice this means that we require an adequate number of observations that are sufficiently similar to a given treated observation to make a good guess about the missing $Y_0$. Failing this we may decide that it is inappropriate to try to make inferences about that particular observation.

How can we assess whether we have sufficient information to proceed? One option would be to check for each variable whether its distribution covers the same range across the two treatment groups, often summarized by a standardized difference in mean. However, it is possible that there is overlap only on the margins of each variable but not in their joint distribution.

A commonly advocated alternative is to focus only on the overlap in the propensity score [18,36]. This is certainly a simpler option and is justified by the theory if you have access to the true propensity score or a reasonably good prediction. However, this strategy places heavy emphasis on the propensity score estimation strategy. If the propensity score specification is incorrect, this can lead to misjudgements about whether or not overlap is satisfied. This problem can be compounded if the propensity score model includes covariates that are not actually confounders, as the additional variables might strongly predict the treatment – and hence degrade propensity score overlap – and yet have no influence on the outcome.

Hill and Su [15] discuss in more detail this idea that if the covariates in our model are a superset of the true confounders, then propensity score based approaches to identifying lack of overlap can lead to overly conservative judgments about which observations need to be excluded from an analysis. In contrast BART-based approaches to identifying observations that lack overlap [15] are geared toward discovering and mitigating threats due to lack of overlap from variables that are most likely to be true confounders. Thus it is more likely to yield estimates that satisfy common *causal* support [15].

We illustrate the approach in Figure 20.7 using a slightly modified version of our original example. In this scenario $14\%$ of the treatment observations (those who used the hyperShoe) are younger than the youngest member of the control group (those who did not use the hyperShoe) and $11\%$ of the control observations (those who did not use the hyperShoe) were older than the oldest member of the treatment group.

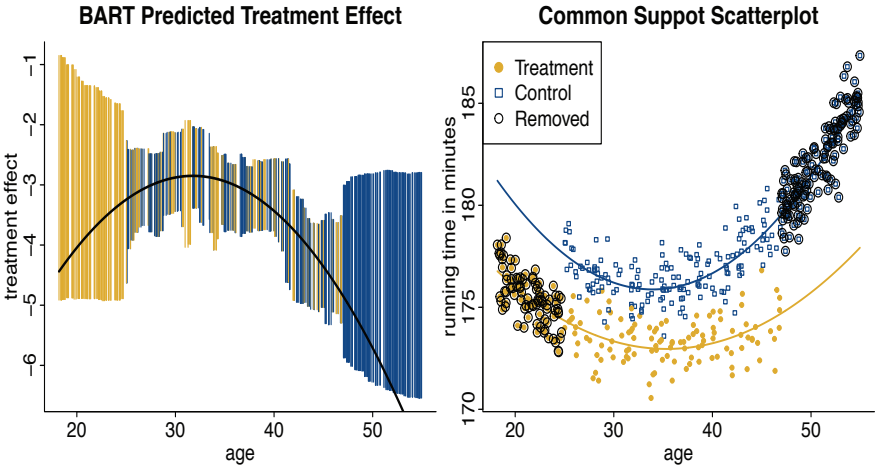


**FIGURE 20.7**
The left panel displays the average treatment effect as it varies with age as a solid black curve. The vertical lines represent 95% uncertainty intervals for the individual level treatment effect for each individual in the sample. The lighter lines correspond to treatment observations. The darker lines correspond to control observations. The right panel displays the response surface and observations again but circles the observations that BART would flag as being at high-risk of lacking common causal support. These assessments are related to the level of uncertainty in treatment effect estimation displayed in the left panel.

The left hand panel of Figure 20.7 displays a curve demonstrating how the true treatment effect varies with age. It also displays 95% uncertainty intervals for the individual-level treatment effects for each of the observations in the sample.[8] The right-hand panel displays the true response surface and observations.

A helpful feature of the BART predictions is that the uncertainty around them expands precipitously once we leave the range of common support. Standard BART measures of overlap [15], implemented in bartCause using the `commonSup.rule` and `commonSup.cut` arguments, would suggest that the circled units in the right-hand plot are sufficiently outside the range of sufficient overlap that we would not trust inferences about them. Therefore, we would discard those observations. The average treatment effect for the remaining observations is $-3.92$ and the BART estimate for those observations is $-3.78$.

The plot on the left also reminds us to be cautious in interpreting individual level treatment effects. While all of the intervals for the observations that would remain in our analysis still cover the true treatment effect (and some of the intervals for the discarded units cover as well!), some of them only just barely cover. Why is that? Because it's an exceptionally difficult inferential task to estimate a treatment effect specific to one covariate value, even when overlap exists in that neighborhood of the covariate space. When we estimate average effects we get to capitalize on the natural bias cancellation that occurs when adding up a bunch of slightly imprecise estimates, which leads to more accuracy for the average effects.

On the other hand this type of individual level prediction is much harder to do with most matching and weighting strategies that don't get to borrow strength across observations in the way

[8]Technically these are intervals of individual level *conditional* average treatment effects. For instance the interval displayed for a person who was 38 years old is just the interval for the average effect for anyone aged 38 in the sample. Formally, we can express each treatment effect as $\mathrm{E}[Y_{1i} - Y_{0i} \mid X_i]$, as distinct from, for example, $Y_{1i} - \mathrm{E}[Y_{0i} \mid X_i]$ (for a treatment observation) or $\mathrm{E}[Y_{1i} \mid X_i] - Y_{0i}$ (for a control observation).

that regression modeling approaches can. Even if we don't want to trust any given individual level prediction or interval we can still capitalize on the fact that they are estimated reasonably well to facilitate better understanding of trends in treatment effect heterogeneity, as is discussed in the next sections.

### 20.6.2 Treatment effect heterogeneity

The BART approach to causal inference, with its combination of flexible modeling embedded in a Bayesian likelihood framework, provides the opportunity for simultaneous inference on individual-level treatment effects as well as any of a variety of average treatment effects (ATE, ATC, ATT, or any subgroup effect defined by measured characteristics).

Hill [28] demonstrated the potential for BART to accurately detect individual-level treatment effects in scenarios where treatment effect heterogeneity is governed by observed covariates. Green and Kern [37] expanded on this to directly address the advantages of using BART to estimate treatment effect heterogeneity. Hahn et al. have since extended the BART framework to explicitly target treatment effects in a way that facilitates more accurate estimation of heterogeneous treatment effects [35]. A similar approach with a more formal decision-theoretic framework was developed by Sivaganesan and co-authors [38].

The BART output also provides a wealth of opportunities for summarizing the information in the posteriors. A simple but useful summary is a "waterfall" plot of point estimates and uncertainty intervals from posterior distributions for the CATE of every sample unit. Typically these plots order the treatment effects estimates and intervals by the magnitude of the treatment effect estimate (e.g., posterior mean or median).

An alternative to waterfall plots is demonstrated by Green and Kern [37] who present partial dependence plots [39] generated by estimating and averaging counterfactual treatment effect functions, which in turn are generated by manipulating potential effect moderators (instead of counterfactual predictions as in typical partial dependence plots). This approach can help us understand how the conditional average treatment effects vary with covariates.

Figure 20.8 provides another possibility. This figure plots data from an extension of our earlier example by considering the treatment effect of hyperShoe moderated by runners age and "mileage," (whether they run a high, moderate or low number of miles per week to prepare for the marathon). In the left panel of the figure we plot posterior distributions of the ATE associated with each of the three mileage groups. While there is some separation in the treatment effect distributions across these groups, there aren't strong differences between them. The right panel further distinguishes treatment effects by age and mileage group. Here we see that once we additionally condition on age we see much stronger differences in treatment effects across groups. Overall, older runners and runners who run more miles per week to prepare receive a greater benefit from the hyperShoe than younger runners and runners who run fewer miles per week.

### 20.6.3 Treatment effect moderation

The ability of BART to identify individual level treatment effects has additional benefits. As we saw in the previous example it allows us to explore what characteristics of observations are associated with variation in treatment effects. This is sometimes referred to as treatment effect moderation.

As another example, in recent work Carnegie et al. [40] explored treatment effect modification in an education example through a variety of graphical summaries. The left panel from Figure 1 of that paper displays a scatter plot of school-specific treatment effects versus a measure of average school-level fixed mindset beliefs. This plot was created to explore whether fixed mindset acted as a moderator of the treatment effect. However, unexpectedly, the plot revealed a cluster of schools with substantially different treatment effects. To better understand these differences a simple regression tree model was fit with estimated treatment effects (output from the BART fit) as the response and

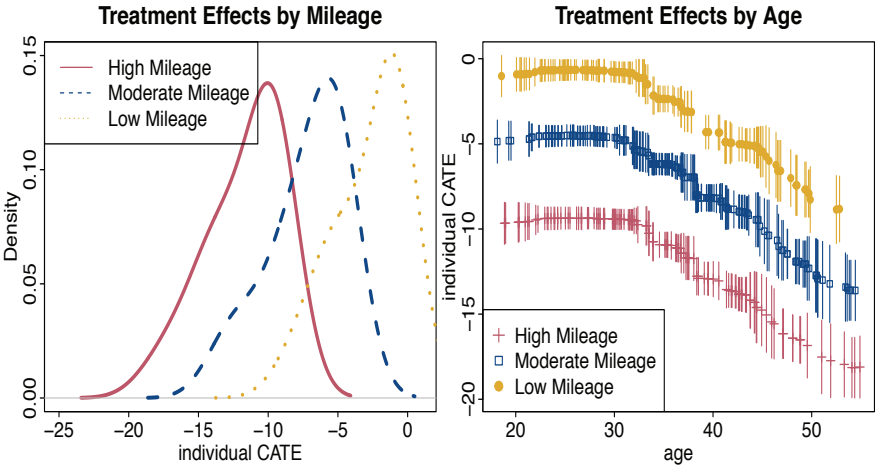


**FIGURE 20.8**
The left panel shows posterior distributions of the ATEs specific to "mileage" groups. While there is separation in the means of these distributions there is still a fair amount of overlap across these distributions. The right panel shows individual-level treatment effects as they vary with respect to levels of both mileage and age. If we hold age constant, we see more distinct separation in the treatment effect posterior distributions across mileage groups.

with the covariates as predictors. This fit revealed that the modifier creating this clustering was a measure of "urbanicity" of the schools. This is a simple and effective technique for discovering moderators and understanding more about treatment effect heterogeneity. The thinkCausal software mentioned previously automates implemention of this strategy (`https://apsta.shinyapps.io/thinkCausal/`).

Causal BART models also yield full posterior distributions over the outcomes under treatment and control. This can facilitate a formal decision theoretic treatment of optimal treatment selection for individuals by selecting the treatment rule that maximizes individual posterior expected utility. Logan et al. [41] use BART in this framework to estimate individualized treatment rules.

We note one caveat about these analyses. We recommend that exploration of treatment effect moderation and estimation of individual-level treatment effects is most plausible in the settings of randomized experiments. All-confounders-measured is a strong assumption in observational settings. Moreover, the overlap assumption is much stronger in observational settings, even if all-confounders-measured is satisfied. Satisfying these at an individual level is more difficult than at a group level. These complications compound in an observation setting whereas in a randomized experiment setting we mostly need to worry only about one of these issues.

### 20.6.4 Generalizability

When treatment effects vary across observations (e.g., students or schools) it is likely that the average treatment effect for a given sample will not be the same as that for another sample. How can we generalize the results from our original sample to a new sample or population that may represent a different distribution of treatment effects? For instance, suppose we find through a randomized experiment conducted in a dozen schools that an intervention was effective for lower performing students but had no impact on other students. If we want to have a sense of how that same intervention might impact a school with a different composition of lower and higher performing students, we would need to reweight (explicitly or implicitly) our estimate to mimic the population of interest.

Of course in more real-world applications treatment effects may vary based on a much larger number of individual-level (e.g. student) and group-level (e.g. school) characteristics. Thus simple reweighting strategies would be more complicated to implement, particularly if the treatment effect modifiers are not known at the outset. A BART approach capitalizes on its ability to estimate individual level effects. If the target population is simply made up a different compositions of the same types of observations that the algorithm was fit to originally,[9] then this task reduces to a prediction problem for a new sample of individuals (or groups) defined by these covariate values. If covariates are available for this new population then we can use BART to generate posterior distributions for both potential outcomes, which in turn can be used to generate posterior distributions for the treatment effect for each person in that new sample and any average effects based on groupings of these people.

If outcomes are additionally available (for instance we know test scores in the absence of an intervention but want to predict test scores (and thus effects) that would occur given exposure to the intervention then the prediction problem is less difficult because only half of the information is missing (the counterfactual outcome). BART-based strategies for generalizing treatment effects were discussed in [43], where BART was demonstrated to have superior performance over propensity score based approaches to generalization in the setting where all confounders are measured and we observe the covariates that modify the treatment effect. Generalization of average treatment effects from one sample or population to another requires an accurate portrayal of the way that treatment effects vary across subgroups. Thus it is no surprise that BART also demonstrates superior performance in this task.

### 20.6.5 Grouped data structures

When observations have a grouping structure (for example cluster randomized trials, repeated measurements, or individuals nested in schools or hospitals), it is important to explicitly include that structure in the treatment and response models. Failure to do so can omit a source of confounding, ignore correlation between observations, and lead to inaccurate measures of uncertainty. To address this, the bartCause package supports a wide variety of methods for grouped data.

The first decision to consider with grouped data is whether to include the grouping variables as so-called "fixed-effects" or "random-effects." These are also known as "unmodeled" or "modeled" parameters.[10] It is beyond the scope of this chapter to discuss the differences between fixed and random effects at length, although it's often the case that fixed effects are more appropriate when the researcher believes that there is unmeasured confounding at the group level. However, fixed effects can also lead to over-parameterization and noisy estimates in small subgroups. Random effects are typically more appropriate when there is a sizeable number of groups, they can be assumed to be independent of the treatment, and there are small groups that would benefit from being "partially pooled" toward the global average. For a longer discussion on fixed versus random effects in the context of causal inference, see this review chapter on multilevel modeling and causal inference [44] as well as research pointing to the potential dangers of using fixed effects for causal effects when all confounders are not measured [11].

Depending on the level of complexity of the grouping structure, bartCause has two supported interfaces. The first is for "varying intercepts," or models with a single grouping factor where the only predictor of interest is group membership itself. In that case, calling `bartc` and passing it the name of a grouping factor in the `group.by` argument will incorporate that variable. Further options include `use.ranef` to indicate the grouping factor should be random or fixed, and `group.effects`,

[9] Roughly speaking this translates into two assumptions: (1) that selection into each of the groups is ignorable with respect to the potential outcomes or the difference between them, and (2) that overlap exists between these groups. For more details see Stuart et al. [42].

[10] It should be noted that unless the foundational assumptions of Section 20.2.3 apply, the term "effect" is a misleading overstatement.

which determines if subgroup averages estimates should calculated as the result. Within the context of our hyperShoe example, suppose that observations were grouped by country, which might serve as a confounder through level of interest in running or funding available through a sports program. To fit a varying intercept model in bartCause and report subgroup average treatment effects:

```
mlm_fit <- bartc(running_time, hyperShoe, age,
                 data = dat, estimand = "att",
                 group.by = country, group.effects = TRUE,
                 seed = 0)
```

The result:

```
> summary(mlm_fit)
fitting treatment model via method 'bart'
fitting response model via method 'bart'
Call: bartCause::bartc(response = running_time, treatment = hyperShoe,
            confounders = age, data = dat, estimand = "att",
            group.by = country, group.effects = TRUE,
            seed = 0)

Causal inference model fit by:
  model.rsp: bart
  model.trt: bart

Treatment effect (population average):
          estimate     sd ci.lower ci.upper   n
country_1   -3.889 0.2615   -4.401   -3.377  42
country_2   -3.707 0.2466   -4.190   -3.223  49
country_3   -3.610 0.2438   -4.088   -3.132  49
country_4   -4.603 0.2709   -5.134   -4.072  42
country_5   -3.605 0.2353   -4.066   -3.144  52
total       -3.857 0.1592   -4.169   -3.545 234
Estimates fit from 464 total observations
95% credible interval calculated by: normal approximation
population TE approximated by: posterior predictive distribution
Result based on 800 posterior samples times 10 chains
```

More complicated grouping structures can be fit using the `parametric` argument of `bartc`. This argument accepts a full parametric equation that is added to the treatment and response models. Multilevel models, including nested or cross effects and varying slopes, are defined as in the lme4 package ( [45]), using a vertical bar notation ( `(var | group)` ). For example, a varying intercept and slope model for our hyperShoe example that allowed a different coefficient for age by country would use the `parametric` argument of `(1 + age | country)`. More recently, extensions of the BART algorithm that accommodate parametric and multilevel structure have been developed in a package called stan4bart, available in R [27,54].

### 20.6.6 Sensitivity to unmeasured confounding

The primary motivation behind using BART (or propensity scores, etc.) for causal inference is to avoid the bias incurred through misspecification of the response surface, ($E[Y(0)|X]$ and $E[Y(1)|X]$). However, the more difficult assumption to relax (in the absence of a controlled randomized or natural experiment) is the assumption that all confounders have been measured, in large part because this assumption is untestable.

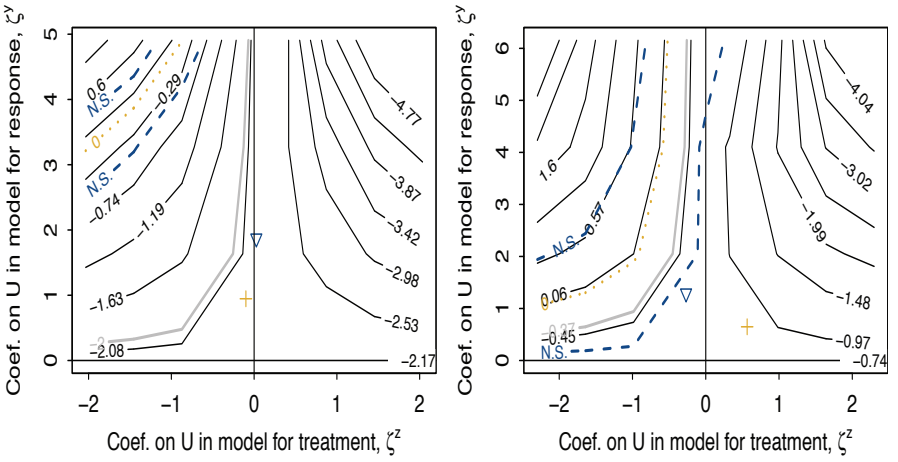


**FIGURE 20.9**
Contour lines corresponding to estimates of the treatment effect corrected for bias indicated by confounding levels at that location on the plot. The left plot displays an example that is not particularly sensitive to unobserved confounding. The right plot corresponds to an example that is more sensitive to unobserved confounding.

One approach to this is to evaluate the sensitivity of a study to potential unmeasured confounding across a variety of assumptions about the potential strength of that confounding. This allows the researcher to understand what level of confounding would be needed to substantively change the estimate of the treatment effect. For instance, what would it take to change the sign of this estimate or drive it to zero?

There is much work in this area but we will briefly focus on work by Hill and co-authors ( [46]), extending earlier work by Carnegie, Harada, and Hill, [47] that allows BART to be incorporated into an existing sensitivity analysis framework. The original approach imposed a strict parametric model with the two parameters for to capture the role of an unobserved confounder; the extension relaxes the assumptions by allowing BART to model the response surface. This results in an easily interpretable framework for testing the potential impact of an unmeasured confounder that also limits the number of modeling assumptions. The performance of this approach was evaluated in a large-scale simulation setting and its usefulness was also demonstrated with high blood pressure data taken from the Third National Health and Nutrition Examination Survey [46].

To illustrate this approach we extended our original example to create two additional scenarios, both of which include an unobserved confounder. We conceive of this variable as a indicator of whether each runner had sponsorship for the race. Sponsorship is positively associated with the probability of having the hyperShoe and negatively associated with the running time (that is if you have sponsorship you are likely to have faster running times on average). In the first scenario the confounding was relatively weak and in the second it was relatively strong. Figure 20.9 displays results.

First consider the plot on the left. Each coordinate in the plot region represents a combination of sensitivity parameters. These parameters reflect the strength of association between our binary unobserved confounder, $U$ (sponsorship), and the treatment ($x$-axis) as well as the strength of the unobserved confounder and the outcome ($y$-axis). In particular we can think of the parameter on the $y$ axis as the difference in means in the outcome between groups with $U = 0$ and $U = 1$ after (non-parametrically) adjusting other the other covariates available; this parameter has been standardized to be represented in standard deviation units (with respect to the outcome variable). The parameter on the $x$-axis represents the coefficient on $U$ in a probit regression model of $Z$ on $U$

and the other covariates. Each contour line in the plot reflects the set of such points (combinations of sensitivity parameters) that would result in a particular (standardized) estimate of the treatment effect, the magnitude of which is displayed on the line. The lighter dashed contour corresponds to a treatment effect estimate of 0 and the darker long-dashed contours show when the estimate would lose statistical significance. Finally, the plus sign and diamond represent that actual coefficients on the other two covariates in the model for these two equations (the triangle presents the estimate with a reversed sign so it is in a quadrant of the space represented by the plot). These help us to provide some context for the size of the coefficients and what might be considered a large magnitude for each of the sensitivity parameters.

Therefore, the left-hand plot, corresponding to the situation where the confounding is weak, accurately reflects that situation. It tells the researcher that a missing confounder would have to have a strong, negative relationship with the treatment and and exceptionally strong relationship with the outcome to drive the treatment effect estimate to zero; the sizes of strengths of those relationships would far exceed those of the observed confounders. On the other hand, the plot on the right displays a situation where the confounding is much stronger. In this case the results are much more sensitive to the unobserved confounder. The treatment effect estimates could be driven to zero with much more moderate associations between $U$ and the treatment and outcome.

## 20.7 Evidence of Performance

Hill [28] first proposed use of BART in causal inference and provided simulation evidence of superior performance relative to simple propensity score approaches. Since then, many other papers have provided additional evidence of the advantages of BART-based methods to causal inference [35, 37, 43, 48, 49]. Although BART has proven to be particularly effective in settings with a single point in time binary treatment variable where all confounders measured holds, in principle any flexible regression model could be used in a similar manner and might have similar properties. The 2016 Causal Inference Data Analysis Challenge [50], associated with the annual Atlantic Causal Inference Conference (ACIC[11]), allowed researchers to submit any of a variety of causal inference algorithms to estimate causal effects across a variety of settings in this type of setting and indeed there were several promising machine-learning based submissions, including BART.[12]

## 20.8 Strengths and Limitations

This section outlines both strengths and limitations of using BART for causal inference. It also highlights directions for ongoing and future work towards ameliorating some of the existing shortcomings.

### 20.8.1 Strengths

BART has three key features that contribute to its strengths: (1) flexible modeling strategy, (2) use of the outcome variable, and (3) the Bayesian framework. We discuss the advantages of a BART approach to causal inference by framing them in terms of their relationship to one or more of these features.

[11] ACIC now stands for American Causal Inference Conference.

[12] The 2017, 2018, and 2019 incarnations of these challenges also demonstrated superior performance for BART. The 2017 results are available as a technical report [51]. 2018 and 2019 results are currently unpublished, but were announced at each of the respective ACIC conferences.

One of the biggest strengths of BART is that it allows for robust estimation of a wide variety of estimands, ranging from average treatment effects, to subgroup effects, to individual-level treatment effects. This versatility results as a combination for the flexible sum-of-trees model and the Bayesian framework which allows us to produce a full posterior distribution for each combination of observation and potential outcome. The ability to produce more robust estimates of individual treatment effects not only allows us to better understand treatment effect heterogeneity but also to explore what covariates moderate treatment effects. The fact that the BART modeling approach incorporates the outcome variable allows for more efficient estimation of these estimands as well.

Capitalizing on the information in the outcome variable also provides BART with a way of implicitly identifying which covariates are true confounders, based on which are most strongly predictive of the outcome. The algorithm then weights the contributions of those variables more strongly. In conjunction with the Bayesian framework which provides a principled strategy for estimating uncertainty, BART can identify observations without sufficient common causal support more easily than many other approaches to causal inference [15].

Additional advantages of the Bayesian framework include the ability to expand the model fairly easily to accommodate extensions. Currently extensions include the stan4bart multilevel BART model [32] and an amalgamation of the BART algorithm into the treatSens sensitivity analysis package, both described above. These are but two of many potential opportunities.

A further strength relates to the high performance of the default settings of most BART implementations. This yields an automated approach which allows the researcher to more easily pre-specify their model which, in turn, limits researcher degrees of freedom and makes the research more easily reproducible. This feature of BART makes it compatible with other "honest" approaches to causal inference because the researcher will not have the opportunity to adjust their model specification as a response to initial looks at treatment effect estimates. While many propensity score approaches also allow for this type of honesty because it is possible to choose a strategy without making use of the outcome variable, these strategies often still allow for many researcher degrees of freedom as the researcher searches for an optimal specification. The full modeling path can be difficult to reproduce for similar reasons.

### 20.8.2 Limitations and potential future directions

There are several limitations to the "vanilla" BART approach to causal inference. One of the most glaring is that it assumes normally distributed, independent, and identically distribution error terms. Extensions of this framework to incorporate group data structures (discussed above) help to weaken the independence assumption. Versions of BART that accommodate heteroskedastic errors address this limitation as well [52].

In addition, several different flavors of BART now exist to address the need to allow for a wider variety of distributional assumptions about the outcome and error term [53]. Refinements of these basic strategies to, for instance, use cross-validation to tune the additional hyperparameters required have been developed as well [46]. More work can be done to relax these assumptions and, ideally, allow the algorithm to automatically detect the right modeling choices.

This approach, like all causal inference approaches, assumes sufficient overlap. While BART has been used to develop some advances in the detection of observations that lack overlap in the confounder space, these approaches are still somewhat ad hoc and could be improved to target specific estimands (such as individual-level treatment effects).

Hahn et al. [35] have pointed out the potential for the most basic implementation of BART for causal inference to induce confounding through the regularization built into the prior. They have created an extension called Bayesian causal forests that provides a promising way to address this problem [35]. They also suggest that an approximate solution is to simply include a reasonable estimate of the propensity score as a covariate in the BART model.

Currently the standard BART approach is mostly useful for studies with a single binary treatment that occurs at one point in time. Many extensions to more complicated settings should be relatively straightforward but don't currently exist. Moreover, the primary implementations are in the R programming language. However stand-alone, user friendly software now exists that provides access to the software without requiring the user to program in R (`https://apsta.shinyapps.io/thinkCausal/`).

Finally, BART approaches to causal inference typically assume that all confounders have been measured. While BART has been incorporated into existing sensitivity analysis approaches [46], this doesn't entirely remove that problem. Use of these approaches requires humility in understanding what conclusions can reliably be drawn and transparency about the assumptions.

## 20.9 Conclusion

This chapter has introduced the basics of a causal inference approach that capitalizes on a Bayesian machine-learning-based algorithm, BART. It explains when and how flexible regression-based approaches to causal inference can be useful and has highlighted some potential advantages of these approaches relative to approaches that focus on data restructuring such as matching and weighting.

To our knowledge BART was the first machine-learning-based approach to causal inference introduced (with scholarly talks starting in 2005 and journal publication in 2011 [28]). However since that time several other machine-learning based approaches to causal inference have also been developed [55–59]. Many of these algorithms have similar desirable features. A distinguishing characteristic of BART is that the flexible model for the response surface is embedded within a Bayesian likelihood framework. This offers advantages with regard to uncertainty quantification, detection of observations that lack common causal support, simultaneous identification of a wide variety of causal estimands, and the ability for reasonably straightforward model extensions to accommodate features such as grouped data structures, sensitivity analysis, and varying distributional assumptions. Both the simple BART-causal implementation and many of the additional features described in this chapter are available in the `bartCause` package in R and the standalone thinkCausal software (`https://apsta.shinyapps.io/thinkCausal/`).

## 20.10 Acknowledgements

This research was supported by funding from the Institute of Education Sciences (R305D200019).

## References

[1] Hugh Chipman, Edward George, and Robert McCulloch. Bayesian ensemble learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

[2] H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298, 2010.

[3] Jennifer Hill, Antonio Linero, and Jared Murray. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7(1):251–278, 2020.

[4] Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and Other Stories*. Cambridge University Press, New York, 2020.

[5] Donald B. Rubin. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58, 1978.

[6] B. S. Barnow, G. G. Cain, and A. S. Goldberger. Issues in the analysis of selectivity bias. In E. Stromsdorfer and G. Farkas, editors, *Evaluation Studies*, volume 5, pages 42–59. Sage, San Francisco, 1980.

[7] Sander Greenland and James M Robins. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419, 1986.

[8] Michael Lechner. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In Michael Lechner and Friedhelm Pfeiffer, editors, *Econometric Evaluation of Labour Market Policies*, volume 13 of *ZEW Economic Studies*, pages 43–58. Physica-Verlag HD, 2001.

[9] Paul R. Rosenbaum. *Observational Studies*. Springer, New York, 2002.

[10] J. Pearl. On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425–432, Catalina Island, CA, 2010. Accessed 02/02/2016.

[11] J. Middleton, M. Scott, R. Diakow, and J. Hill. Bias amplification and bias unmasking. *Political Analysis*, 24:307–323, 2016.

[12] P. Steiner and Y. Kim. The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of Causal Inference*, 4, 2016.

[13] P. Ding and L. Miratrix. To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference*, 3:41–57, 2014.

[14] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

[15] Jennifer Hill and Yu-Sung Su. Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Annals of Applied Statistics*, 7:1386–1420, 2013.

[16] Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008. The regression discontinuity design: Theory and applications.

[17] Sebastian Calonico, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. Rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, 17(2):372–404, 2017.

[18] Guido Imbens and Donald Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, New York, 2015.

[19] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.

[20] Luke Keele. The statistics of causal inference: A view from political methodology. *Political Analysis*, 23:313–35, 2015.

[21] B. Lara, J. Salinero, and J. Del Coso. The relationship between age and running time in elite marathoners is u-shaped. *Age (Dordr)*, 36(2):1003–1008, 2014.

[22] Niklas Lehto. Effects of age on marathon finishing time among male amateur runners in stockholm marathon 1979–2014. *Journal of Sport and Health Science*, 5(3):349–354, 2016.

[23] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2 edition, 2003.

[24] Luke Tierney. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701 – 1728, 1994.

[25] George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[26] Vincent Dorie. *stan4bart: Bayesian Additive Regression Trees with Stan-Sampled Parametric Extensions*, 2021. R package version 0.0-1.

[27] Vincent Dorie. *stan4bart: Bayesian Additive Regression Trees with Stan-Sampled Parametric Extensions*, 2021. R package version 0.0-1.

[28] Jennifer Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[29] Hugh Chipman and Robert McCulloch. *BayesTree: Bayesian Additive Regression Trees*, 2016. R package version 0.3-1.4.

[30] Adam Kapelner and Justin Bleich. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016.

[31] Vincent Dorie, Hugh Chipman, and Robert McCulloch. *dbarts: Discerete Bayesian Additive Regression Trees Sampler*, 2014. R package version 0.8-5.

[32] Bereket Kindo. *mpbart: Multinomial Probit Bayesian Additive Regression Trees*, 2016. R package version 0.2.

[33] Belinda Hernandez. *bartBMA: Bayesian Additive Regression Trees Using Bayesian Model Averaging*, 2020. R package version 1.0.

[34] Robert McCulloch, Matthew Pratola, and Hugh Chipman. *rbart: Bayesian Trees for Conditional Mean and Variance*, 2019. R package version 1.0.

[35] P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[36] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

[37] Holger L Kern, Elizabeth A Stuart, Jennifer Hill, and Donald P Green. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, pages 1–25, 2016.

[38] Siva Sivaganesan, Peter Müller, and Bin Huang. Subgroup finding via bayesian additive regression trees. *Statistics in Medicine*, 36(15):2391–2403, 2017.

[39] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[40] N. Carnegie, V. Dorie, and J. Hill. Examining treatment effect heterogeneity using BART. *Observational Studies*, 5:52–70, 2019.

[41] Brent R Logan, Rodney Sparapani, Robert E McCulloch, and Purushottam W Laud. Decision making and uncertainty quantification for individualized treatments using bayesian additive regression trees. *Statistical Methods in Medical Research*, page 0962280217746191, 2017.

[42] Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society A*, 174(2):369–386, 2011.

[43] Holger L Kern, Elizabeth A Stuart, Jennifer Hill, and Donald P Green. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, pages 1–25, 2016.

[44] Jennifer Hill. Multilevel models and causal inference. In Marc A. Scott, Jeffrey S. Simonoff, and Brian D. Marx, editors, *The SAGE Handbook of Multilevel Modeling*. Sage Publications Ltd, 2013.

[45] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[46] Vincent Dorie, Masataka Harada, Nicole Carnegie, and Jennifer Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine*, 35(20):3453–3470, 2016.

[47] Nicole Bohme Carnegie, Masataka Harada, and Jennifer Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9:395–420, 2016.

[48] Jennifer L. Hill, Christopher Weiss, and Fuhua Zhai. Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, 46:477–513, 2011.

[49] T Wendling, K Jung, A Callahan, A Schuler, NH Shah, and B Gallego. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 2018.

[50] V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

[51] P. Richard Hahn, Vincent Dorie, and Jared S. Murray. Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017. *arXiv e-prints*, page arXiv:1905.09515, May 2019.

[52] M. T. Pratola, H. A. Chipman, E. I. George, and R. E. McCulloch. Heteroscedastic bart via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, 29(2):405–417, 2020.

[53] Jared S. Murray. Log-linear bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association*, 116(534):756–769, 2021.

[54] Vincent Dorie, George Perrett, Jennifer L. Hill, and Benjamin Goodrich. Stan and bart for causal inference: Estimating heterogeneous treatment effects using the power of stan and the flexibility of machine learning. *Entropy*, 24(12):1782, Dec 2022.

[55] The H2O.ai team. *h2o: R Interface for H2O*, 2016. R package version 3.10.0.10.

[56] Erin LeDell. *h2oEnsemble: H2O Ensemble Learning*, 2016. R package version 0.1.8.

[57] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[58] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, Feb 2019.

[59] Cheng Ju, Susan Gruber, Samuel D Lendle, Antoine Chambaz, Jessica M Franklin, Richard Wyss, Sebastian Schneeweiss, and Mark J van der Laan. Scalable collaborative targeted learning for high-dimensional data. *Statistical Methods in Medical Research*, 28(2):532–554, 2019. PMID: 28936917.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com